

AD-A130 804

INDEXING FOR INFORMATION RETRIEVAL IN THE INTEGRATED

1/1

LIBRARY SYSTEM(U) MITRE CORP MCLEAN VA METREK DIV

R A DUNCAN 14 SEP 79 WP-79W00620 LHNBCB-CR-81-10

UNCLASSIFIED

N01-LM-8-4720

F/G 5/2

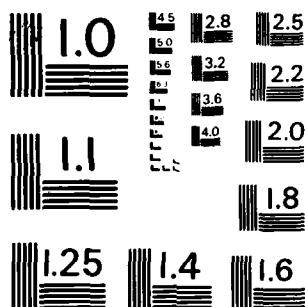
NL

END

DATE

FILED

DTIC

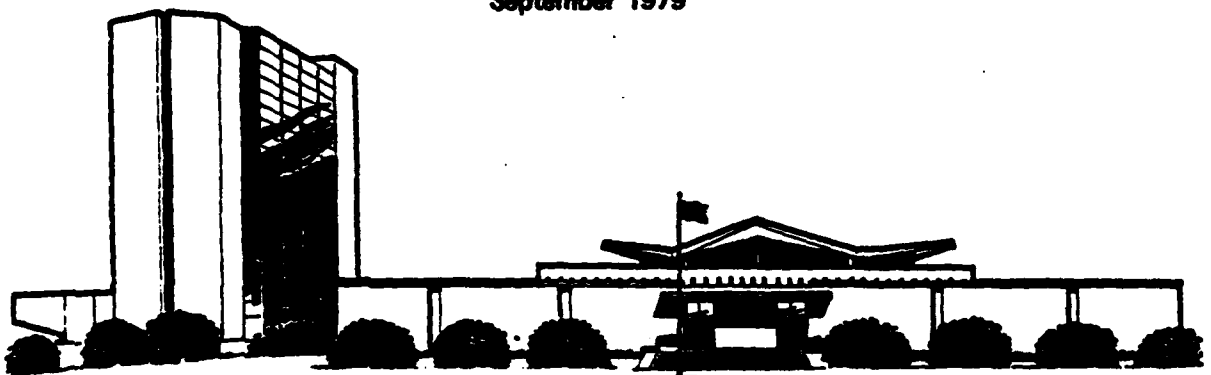


MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AD A130804

INDEXING FOR INFORMATION RETRIEVAL IN THE INTEGRATED LIBRARY SYSTEM

September 1979



DTIC FILE COPY

This document has been approved
for public release and sale; its
distribution is unlimited.

DTIC
SELECTED
JUL 19 1983
A

59272-101

REPORT DOCUMENTATION PAGE		1. REPORT NO. LHNCBC CR 81-10	2.	3. Recipient's Accession No.
4. Title and Subtitle Indexing for Information Retrieval in the Integrated Library System				5. Report Date September 14, 1979
7. Author(s) Duncan, Roger A., The MITRE Corporation				6. Performing Organization Rept. No. WP-79W00620 ✓
9. Performing Organization Name and Address The MITRE Corporation Metrek Division 1820 Dolley Madison Boulevard McLean, VA 22102				10. Project/Task/Work Unit No. 1801E
				11. Contract(C) or Grant(G) No. (C) NO1-LM-8-4720 (G)
12. Sponsoring Organization Name and Address Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Public Health Service, Department of Health and Human Services				13. Type of Report & Period Covered Contractor Report
15. Supplementary Notes				14.
16. Abstract (Limit: 200 words) This report reviews alternative indexing schemes for retrieval of bibliographic information in several current automated systems, considers others proposed in meetings held among librarians and developers of the Integrated Library System, and recommends indexing schemes which will serve both librarians and library patrons.				
17. Document Analysis a. Descriptors b. Identifiers/Open-Ended Terms Indexing Library Automation Library Systems c. COSATI Field/Group				
18. Availability Statement: Unclassified - Unlimited		19. Security Class (This Report) Unclassified		20. No. of Pages 20
		20. Security Class (This Page) Unlimited		22. Price

(See ANSI-Z39.18)

See Instructions on Reverse


OPTIONAL FORM 272 (4-77)
(Formerly NTIS-35)
Department of Commerce

TABLE OF CONTENTS

	<u>Page</u>
1. INTRODUCTION	1-1
1.1 Background	1-1
1.2 Purpose and Scope	1-2
2. INDEXING SCHEMES IN OTHER SYSTEMS	2-1
2.1 OCLC, Inc.	2-1
2.2 Washington Library Network	2-2
2.3 Research Libraries Information Network	2-3
2.4 Ohio State University Libraries Automated Circulation	2-3
2.5 University of Toronto Library Automation Systems	2-4
2.6 Summary	2-4
3. ALTERNATIVES CONSIDERED FOR THE INTEGRATED LIBRARY SYSTEM	3-1
3.1 Subject Headings	3-1
3.2 Series	3-3
4. CONCLUSIONS AND RECOMMENDATIONS	4-1
BIBLIOGRAPHY	B-1
APPENDIX - INFORMATION ON OTHER SYSTEMS	A-1
DISTRIBUTION LIST	D-1

LIST OF TABLES

TABLE I	SUMMARY OF SYSTEM SEARCH FEATURES	2-5
---------	-----------------------------------	-----

<p><i>Per Mr. Charles Gould</i></p> <p><i>Letter on file</i></p> <p style="text-align: center; font-size: 2em;">A</p>	<div style="text-align: center;">  </div> <p style="text-align: center;">111</p>
---	--

1. INTRODUCTION

1.1 Background

The Computer Technology Branch of the Lister Hill National Center for Biomedical Communication, National Library of Medicine, has been developing the Integrated Library System since 1977. The system is intended to offer a "user-cordial interface"¹ to facilitate entry of and access to bibliographic information, whether the user is experienced or unfamiliar with the system's features and internal structure. Furthermore, the system is being designed for a library to tailor it to its own needs. Thus, the system is largely parameter- and table-driven, so that any library which chooses to use the system can establish, e.g., its own limits on the number of items a given class of patron is permitted to borrow.

At present, the circulation module of the system has been developed. Other modules, such as acquisition, are in various stages of design and development.

A fundamental issue in design of the system both for cataloging bibliographic material and efficient and effective retrieval of information on these materials, is the manner in which information may be indexed. For example, a patron who wishes to retrieve information on a bibliographic item may know only a few terms in the title of the work. If the system is indexed in such a way that only knowledge of the full title will permit retrieval, the patron will not be able to gain access to the desired information. Even if retrieval is possible knowing the first x characters of the title's first "significant" term, the first y characters of the next, and the first z of the next, the patron may not know enough terms of the title or their sequence to complete a successful retrieval.

¹Goldstein, C. M. and Ford, W. H., "The user-cordial interface," On-line Review, 2, 3, 1978, pp. 279-275.

1.2 Purpose and Scope

The purpose of this paper is to explore alternative indexing schemes and recommend from among them those which conform to the design philosophy of the Integrated Library System. The paper discusses indexing schemes used by other automated library systems which are prominent in the library community today. This document also presents alternative approaches considered in two meetings of librarians recently held at the National Library of Medicine. The pros and cons of the alternatives are considered and conclusions are drawn and recommendations are made.

This report is not intended to provide an exhaustive review of indexing schemes but rather address those which are considered both reasonable for users and attainable at acceptable cost.

2. INDEXING SCHEMES IN OTHER SYSTEMS

The indexing features of five major automated library systems were analyzed for this report. The systems studied were those of

- OCLC, Inc.,
- the Washington Library Network (WLN),
- the Research Libraries Information Network (RL N) (formerly known as BALLOTS),
- the Ohio State University Libraries Automated Circulation System, and
- the University of Toronto Library Automation Systems (UTLAS).

Each is treated in turn in the following subsections. The Appendix contains further details on the indexing features of these systems.

2.1 OCLC, Inc.

OCLC, Inc. (formerly the Ohio College Library Center) is a bibliographic utility which takes MARC-formatted bibliographic record information from tapes supplied by the Library of Congress (LC), reformats the information, and offers it to subscriber libraries and library networks online, in remote batch form, or in the form of tapes or cards for their catalogs. Retrieval may be made on author key, title key, combination author and title key, LC card number, International Standard Book Number (ISBN), International Standard Serial Number (ISSN) (in the case of journals, magazines, newspapers, and the like), CODEN (a five-character code for the title of a serial), and OCLC control number. The author, title, and author-title keys are set up in the following formats, where the numbers correspond to the first digits in each major term. In the case of an author, the sequence is surname, first name, middle name. In the case of titles, the sequence is major terms as they appear in the title.

AUTHOR: 4,3,1; 4,2,1; 4,1,1; 4,3; 4,3; 4,1,
TITLE: 3,2,1,1; 3,1,1,1; 3,1,2,1
AUTHOR-TITLE: 4,4; 4,3; 3,4; 3,3.

2.2 Washington Library Network

WLN offers access via author, title, and subject, although in the last case, the system user must know the proper LC subject heading. Access is possible at two levels, in the authority file and in the bibliographic file. In the authority file, one can access via

AUTHOR	generic, personal, corporate, meeting or conference (full text and keyword)
TITLE	uniform title heading (e.g., BIBLE) (full text and keyword)
SERIES	generic, personal, corporate, meeting or conference, uniform title heading, or series title (traced series only) (full text and keyword)
SUBJECT	generic, personal, corporate, meeting or conference, uniform title heading, topical and geographic subjects (full text and keyword).

In the bibliographic file, one can enter via

AUTHOR	generic and via corporate/conference keywords
TITLE	uniform title
SERIES	title
SUBJECT	generic and via corporate/conference keywords

In addition, there are other indexes to the bibliographic file: LC card number, ISBN, ISSN and keywords from title or subject in any order.

2.3 Research Libraries Information Network

RLIN is the system of another bibliographic utility (also the name of the utility) formed by the Research Library Group. The system offers access to its bibliographic files via personal, corporate or conference author (full or truncated form, in normal or inverted sequence), title keyword(s) (in any sequence, and truncated if desired), and subject heading (full or truncated). The user must know the correct heading (e.g., by referring to LC's "Red Book," the authority for subject headings). In addition, the user can also access information via call number, LC card number and local identification number. Boolean search is also possible in which, e.g., author. and .title, are linked.

2.4 Ohio State University Libraries Automated Circulation System

The Ohio State University's circulation system is known as the Library Control System (LCS). It offers general and detailed search capabilities. The general search retrieves a single line of information containing main entry, title edition statement, and publication date. The general search can be conducted via author-title keys or subject (the latter is only available for the past one-and-one-half year's worth of system entries). An author-title search key is composed of the first four characters of the author's surname and the first five characters of the first significant title word. A subject search requires that the user key in an entire valid LC subject heading.

The LCS detailed search yields all data in the main file on an item and the circulation status of all copies. The search may be made via

AUTHOR: 6,3

TITLE: 4,5

as well as by call number, title number (as stored in LCS), and display line number obtained from the general search.

2.5 University of Toronto Library Automation Systems

The University of Toronto's system (UTLAS) offers what is known as "extended browsable access" to both authority and bibliographic files in its shared cataloging system. Access is possible to both files by author, title (up to first forty characters), and subject, by LC card number, ISBN, ISSN, NLM control number (for CATLINE items), and special control numbers. All terms can be right truncated to any number of characters. Any two search keys may be combined in Boolean (AND, OR or NOT) search.

The keyword index is being loaded at the present. When this feature is available, there will be keyword access to the bibliographic files by title words and corporate and conference words. The authority files will be accessible by the same keywords plus subject heading words. Searches can also be conducted by specific field (e.g., personal author), if known. Matching on searches is insensitive to diacritics, punctuation, and capitalization.

UTLAS also offers a minicomputer-based online catalog and circulation system with essentially the same features. Searching can also be conducted as through a dictionary or via the specific fields.

2.6 Summary

The information in the above subsections is summarized in Table 2-1. The reader is referred to the preceding sections and the Appendix for greater detail.

TABLE I. SUMMARY OF SYSTEM SEARCH FEATURES

SEARCH FEATURES	SYSTEM	OCLC, Inc.	WASHINGTON LIBRARY NETWORK	RESEARCH LIBRARIES INFORMATION NETWORK	OHIO STATE UNIVERSITY LIBRARIES LCS	UNIVERSITY OF TORONTO LIBRARY AUTOMATION SYSTEMS
ACCESS POINTS						
Author (Generic)			A&B			A&B
Personal Author			A	B		
Corporate/Conference Author			A	B		
Corporate/Conference Keywords			A&B	B,NS		
Author Truncation	ST			B,AT	ST	A&B,AT
Author-Title Truncation	ST				ST	
Title (Uniform)			A			A&B; 240 chars.
Title Keywords			B,NS	B,NS		(A&B)
Title Truncation	S,ST			B,AT	ST	
Subject (Generic)			A&B	B	✓	A&B
Personal Name Subject			A			
Corporate/Conference Name			A			
Subject						
Corporate/Conference Name			A&B,NS			(A&B)
Subject Keywords						(A&B)
Subject Keywords						
Uniform Title Subject			A			
Topical Subject			A			
Geographic Subject			A			
Subject Truncation				B,AT		
Series (Generic)			A			
Personal Name Series			A			
Corporate/Conference Name			A			
Series						
Corporate/Conference Name			A			
Series Keywords						
Title Series			A&B			
Series Keywords						
ISBN		✓	B			A&B
ISSN		✓	B			A&B
CODEN		✓				
System Record ID		✓	B	B	✓	A&B
LC Card Number		✓	Same	B		A&B
Call Number				B	✓	
Other						NLM Control Number
Boolean Search				B		A&B
COMMENTS					General and detailed search modes	Boolean search on any two access points

KEY

A - Authority File
B - Bibliographic File
S - In Sequence
NS - Any Sequence

ST - Specific Right Truncation
AT - Any Right Truncation
() - Future Capability

3. ALTERNATIVES CONSIDERED FOR THE INTEGRATED LIBRARY SYSTEM

Two meetings were convened in August at the National Library of Medicine to consider alternative approaches to indexing with particular attention being devoted to subject headings and series.

3.1 Subject Headings

The problems identified with subject headings were:

- (1) providing a user-cordial interface to the system for librarians and patrons should obviate the need for referral to a printed thesaurus (such as LC's subject authority, Library of Congress: Subject Headings;
- (2) the OCLC fields containing subject information, the 6XX series, may have multiple interpretations;
- (3) the subfields of the 6XX series may also have multiple interpretations and may have multiple occurrences (with different content) of a given subfield for an item.

Typically, as noted for the other systems discussed in Section 2, a user who wishes to access information on an item by subject heading must know either the valid LC subject terms or at least a correct right truncation of one or more of them. The library patron may not realize that "dogs" is a valid subject but that "dog" is not.

An example of multiple interpretations of 6XX fields is the term "Russia" for field 610 (Subject Added Entry - Corporate Name) and for field 651 (Subject Added Entry - Geographic Name).^{*} In both cases, Russia is a place name, but depending on associated indicators and subfields, could refer to White Russia or Soviet Russia. Thus, retrieving on the subject term "Russia," items cataloged

^{*}The author is indebted to Linda Proudfoot of the Department of Justice Library for these examples which were presented at one of the meetings held at NLN.

under 610 and 651 would be retrieved, but without more qualifying information, the search would yield items from several historical periods. If the user retrieved under the corporate name of "Russia," he might not retrieve all items on the subject, since some would be stored under geographic name.

An example of multiple interpretation and occurrences of subfields is for the 650 field of "Vietnamese Conflict, 1961-1975." A possible x subfield (general subdivision) is "Campaigns," while there could be several z subfields (place subdivision), for, e.g., "United States," "Vietnam," and "Cambodia." Furthermore, a place subdivision could have a further place or general subdivision, such that "Vietnam" as a place subdivision could have a further place subdivision of "Son Tay" or a general subdivision of "Tet."

Thus, retrieving on the subject term "Vietnam" as a full word would yield items stored under the 610 and 651 fields, as a truncated term could also yield items stored under the 650 field, and could yield from the subfields as well. A sophisticated user retrieving on "Vietnam" at the 651 field level might not retrieve other relevant items cataloged under 610 or at subfield levels.

The ideas considered at the NLM meetings to address these issues were:

- (1) in lieu of using a printed subject thesaurus, offer the ILS user a menu on the CRT screen;
- (2) offer subject heading search in any order of the words of the heading, with any number of leading characters (i.e., with right truncation);
- (3) use a stop list for common terms (i.e., do not search on such terms as "United States;");
- (4) build the subject heading organizational structure as a hierarchy but do not require the user to know the hierarchy;

- (5) provide free text search among subjects;
- (6) offer Boolean search to link subjects (for narrowing a search);
- (7) go from subject search to media for further restriction (e.g., only consider films).

3.2 Series

Series, whether in the form of monographs, journals, conference proceedings, or other serial publications, present special indexing problems and, again, the user-cordial interface is a fundamental concern. Among the problems are:

- (1) a series may have a series title (e.g., McGraw-Hill Series in Systems Science) as well as a different title for each work in the series;
- (2) the series may be associated with an individual (e.g., Thomas McKillop memorial lectures in public administration), while individual works in the series are written or edited by others;
- (3) volume numbers in a series may be assigned out of the sequence in which the works are published (e.g., volume IV may appear in 1977 and volume II in 1981); volume numbers can be and are also reassigned within a series;
- (4) a series may contain a subseries which is traced (i.e., indexed) differently than the main series (the main series is found in the 4XX fields of the OCLC record format, while the subseries information is found in the 8XX fields);
- (5) a series may have been issued many years ago, and selected individual works in the series may be reissued in more recent times but not as part of a series;
- (6) a series may be known by more than one name (e.g., McGraw-Hill Systems Science Series and McGraw-Hill Series in Systems Science): one form may be in the 4XX, the other in the 8XX fields; furthermore, different works in a series may list the series statement differently;

- (7) different kinds of libraries may choose to store series information in the 4XX field which would, by conventions in other kinds of libraries, be stored in the 8XX field;
- (8) frequently library patrons remember a series title without its associated organization (i.e., as found in the 440 field, when the 490 field is the one to look at for the best retrieval of information on a series); patrons may also remember a popular version of a series title;
- (9) there is no complete authority file for series as there is for author and subject.

The issues cited, together with the desire to produce an ILS which supports both patrons and librarians, offer a challenge to the system designers. The patron should be able to retrieve information on a series or a work in a series without knowing whether his "title" is the "correct" series statement, the title of the work rather than the series, or another (perhaps colloquial) title. The librarian should be able to search either as a patron or using his or her knowledge of the OCLC fields as used in his or her institution.

The ideas raised in the meetings at NLM to treat these issues were:

- (1) offer the user the opportunity to search on keywords in title (so that the precise title doesn't have to be known);
- (2) use Boolean linking of title keywords for access to series information;
- (3) allow search on all series, whether traced (in the current card catalog) or not;
- (4) offer "search only 4XX" capability so that the user does not have to know if he wants data in field 410, 411, etc.;
- (5) allow search for author regardless of which field the name might be found in (110, 410, 810, etc.);
- (6) while offering the patron a generic search capability, provide the librarian access by OCLC field for more targeted retrieval;

- (7) offer a kind of authority file such that, if the user enters a non-standard term or title for a bibliographic search, ILS can translate the non-standard to the authority term or title for retrieval;
- (8) when a large list of items is retrieved on a query (e.g., seven volumes of a series), present the list on the series in order of most recent first (since the user frequently wants the most current item and may not even be aware of earlier issues);
- (9) provide synonyms in the system such that the user can link, e.g., World Health Organization and WHO, to a title provided by a patron to yield Series Statement-Corporate Body, even if patron doesn't know the desired item is part of a series (e.g., World Health Organization Series on Biochemistry);
- (10) offer access via document number, particularly for technical patrons; *T.R.L. #*
- (11) offer ability to order retrievals by call number, by alphabetic sequence of author and title for producing bibliographies and selective dissemination of information.

4. CONCLUSIONS AND RECOMMENDATIONS

In broad terms, the conclusions which are drawn from a review of other systems' search capabilities and points raised by individuals in the library community are as follows.

- (1) The unsophisticated user (i.e., one who does not know the bibliographic "hierarchy" embodied in the OCLC variable field structure) should be able to retrieve bibliographic information by subject without having to refer to a subject authority volume.
- (2) The unsophisticated user should be able to retrieve bibliographic information on series and works in a series (e.g., via author or title) without knowing the hierarchy; however, sophisticated users should be able to direct their search using the hierarchy.
- (3) Searches should be possible using keywords and Boolean operators to link them in lieu of, e.g., knowing a full, correct series statement or subject heading.

Based on the state-of-the-art in bibliographic systems today and concerns expressed by members of the library community who work with patrons, the following recommendations are made.

RECOMMENDATION 1:

Index subject headings such that information is retrievable by

- subject "synonym" (on-line lookup in a thesaurus)
- subject keyword
- right-truncated subject heading
- Boolean combination of subject keywords (linked by AND, OR, and NOT)
- specific OCLC field and subfield.

RECOMMENDATION 2:

Devise an on-line thesaurus to map series statements as they are popularly known to the alternative forms accepted officially (490 and 810).

RECOMMENDATION 3:

Index series such that information is retrievable by:

- correct series statement (whether identified as a series statement or not)
- by series statement keyword (whether identified as series or not) (e.g., author series, title series, corporate or conference series)
- by right truncated keyword
- by Boolean combination of series statement keywords linked by AND, OR and NOT (whether identified as series or not)
- by specific 4XX and 8XX field and subfield.

RECOMMENDATION 4:

Build a stop list, using one or more existing stop lists for other systems (e.g., WLN), for series.

BIBLIOGRAPHY

- Anttila, E., "Authorities at UTLAS: Concepts and Facilities," in Furoya, N.Y., ed.: What's in a name? Control of catalog records through automated authority files. Toronto: University of Toronto Press, 1978.
- Cain, J., "The UTLAS Authority System," in Furoya, N.Y., ed.: What's in a name? Control of catalog records through automated authority files. Toronto: University of Toronto Press, 1978.
- Calk, J., "On-Line Authority Control in the Washington Library Network," in Furoya, N.Y., ed.: What's in a name? Control of catalog records through automated authority files. Toronto: University of Toronto Press, 1978.
- Campbell, B., "The UTLAS CIRC System," University of Toronto Library Automation Systems, Toronto, Ontario.
- Guthrie, G. D., "An On-Line Remote Catalog Access and Circulation Control System," in Proceedings of the American Society for Information Science, 8, 1971, pp. 305-309.
- Proudfoot, L. G., "Searching the OCLC Data Base," The Army Library, Pentagon, Washington, D.C., June 1978.
- Veaner, B., "BALLOTS - The View from Technical Services," Library Resources and Technical Services, 21, 2, Spring 1977, pp. 127-146.
- Velasquez, H., and Anttila, E., "Inter-Library Searching Through an On-Line Catalogue," University of Toronto Library Automation Systems, Toronto, Ontario.

DATE
ILME